

## Robust Machine Learning Applied to Terascale Astronomical Datasets

Nicholas M. Ball, Robert J. Brunner, Adam D. Myers

*Department of Astronomy and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

**Abstract.** We present recent results from the Laboratory for Cosmological Data Mining<sup>1</sup> at the National Center for Supercomputing Applications (NCSA) to provide robust classifications and photometric redshifts for objects in the terascale-class Sloan Digital Sky Survey (SDSS). Through a combination of machine learning in the form of decision trees, k-nearest neighbor, and genetic algorithms, the use of supercomputing resources at NCSA, and the cyberenvironment Data-to-Knowledge, we are able to provide improved classifications for over 100 million objects in the SDSS, improved photometric redshifts, and a full exploitation of the powerful k-nearest neighbor algorithm. This work is the first to apply the full power of these algorithms to contemporary terascale astronomical datasets, and the improvement over existing results is demonstrable. We discuss issues that we have encountered in dealing with data on the terascale, and possible solutions that can be implemented to deal with upcoming petascale datasets.

### 1. Introduction

We summarize work carried out as part of the Laboratory for Cosmological Data Mining, a partnership between the Department of Astronomy and the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign (UIUC), in collaboration with the Automated Learning Group (ALG) at NCSA, and the Illinois Genetic Algorithms Laboratory at UIUC. This combination of expertise allows us to apply the full power of machine learning to contemporary terascale astronomical datasets.

### 2. Data

The SDSS (York et al. 2000) is a project to map  $\pi$  steradians of the Northern Galactic Cap in five broad optical photometric bands,  $u$ ,  $g$ ,  $r$ ,  $i$ , and  $z$ . The Third Data Release of the survey consists of 142,705,734 unique objects, of which 528,640 have spectra. The Fifth Data Release is a superset of DR3, approximately one and a half times the size, consisting of 9.0 TB of images and 1.8 TB of flat files in FITS format. We utilize the objects with spectra as training sets and perform blind tests on subsets of the spectra. The training features are the four colors  $u - g$  through  $i - z$  in the four magnitude types

---

<sup>1</sup><http://lcdm.astro.uiuc.edu>

measured by the SDSS. We provide classifications for the full DR3 sample of 143 million objects.

### 3. Computing Environment

The machine learning algorithms are implemented within the framework of the Data-to-Knowledge toolkit (D2K; Welge et al. 1999), developed and maintained by the ALG at NCSA. This allows the straightforward implementation of numerous Java<sup>TM</sup> modules which automate the stages of the data mining and learning process. The machine learning algorithms available include decision tree,  $k$ -nearest neighbor, artificial neural network, support vector machine, unsupervised clustering, and rule association. For the terascale datasets in use here, our implementation includes enhanced versions of the standard modules which stream data of fixed type, for example single-precision floating point.

The algorithms are run on the Xeon Linux Cluster *tungsten* at NCSA, as part of a peer-reviewed, nationally allocated, LRAC allocation to the LCDM project on this and other machines, renewed over multiple years. Tungsten is composed of 1280 compute nodes. Each node is a Dell PowerEdge 1750 server running Red Hat Linux with two Intel Xeon 3.2 GHz processors, 3 GB of memory, a peak double-precision performance of 6.4 Gflops, and 70 GB of scratch disk space. A further 59 TB of general scratch space is available, and the system is connected via FTP interface to the 5 PB UniTree DiskXtender mass storage system, and the nodes to each other via Myrinet.

### 4. Classification

Using decision trees, we are able (Ball et al. 2006) to assign the probabilities  $P(\text{galaxy, star, neither-star-nor-galaxy})$  to each of the 143 million objects in the SDSS DR3. This enables one to, for example, either emphasize completeness (the fraction of the true number of the target object correctly identified), or efficiency (the fraction of the objects assigned a given type that are correct) in subsamples, both of which have important scientific uses.

### 5. Photometric Redshifts

The left-hand panel of Figure 1 shows a result typical until recently for photometric versus spectroscopic redshift for quasars, here generated as a blind test on 11,149 SDSS DR5 quasars using a single nearest-neighbor model. Most objects lie close to the ideal diagonal line, but there are regions of ‘catastrophic’ failure, in which the photometric redshift assigned is completely incorrect. The kNN enables us to significantly reduce the instance of these catastrophics, as shown in Ball et al. (2007), where the RMS deviation between the photometric and spectroscopic redshifts is reduced from 0.46 as shown to 0.35.

Because every magnitude in the SDSS has an associated error, one can also perturb the testing set numerous times according to the errors on the input features, and use the resulting variation to generate full probability density functions (PDFs) in redshift. The advantage of this is that the errors on the

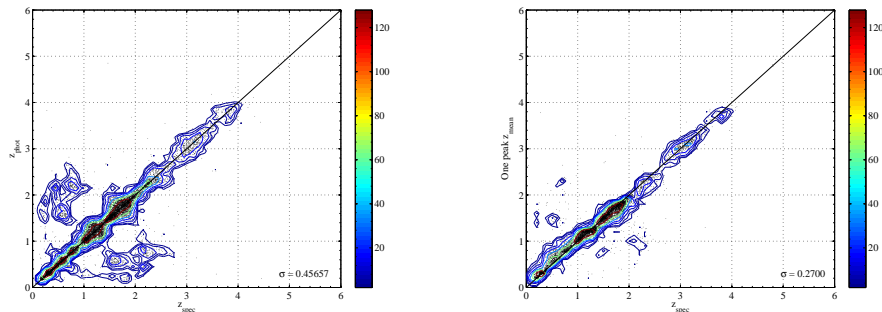


Figure 1. Spectroscopic versus photometric for SDSS DR5 quasars. The left-hand panel shows a result typical of the literature until recently. The right-hand panel shows the result of using machine learning to assign probability density functions then taking the subset with a single peak in probability.

input features are taken into account when assigning the output value. Taking the mean value from each PDF gives a similar RMS dispersion to the 0.35 result, however, for the subset of quasars which have a single PDF peak, the dispersion is further reduced to 0.27, with very few remaining outliers. This is shown in the right-hand panel of Figure 1.

## 6. Discussion

Given the petascale datasets planned for the next decade, it is vitally important that contemporary data mining can be carried out successfully on this scale. In turn, this requires robust techniques on the terascale. We encountered numerous issues that were relevant to realizing this goal:

- Because we are using tens to hundreds of parallel nodes and streaming many GB of data, D2K must be invoked via batch script, negating the advantages of its GUI interface. The resulting lack of an integrated cyberenvironment results in batch scripts that contain many tens of settings, manually set file locations and commands, making them prone to error.
- Job submission is inflexible, subject to fixed wallclock times and numbers of nodes, unpredictable queuing times and no recourse if a job fails due a bug or hardware problem.
- The large datasets must be stored on the Unitree mass storage system, which is occasionally subject to outages in access or significant wait times. In combination with the queuing system for batch jobs described above, this can make new scripts time-consuming to debug.
- There is no way in which to fully explore the huge parameter space (more than  $10^{15}$  combinations of settings for decision trees) of the machine learning algorithms. Genetic algorithms were used to optimize the training features, and could be used similarly to optimize the algorithm.
- The present lack of fainter training data forces us to extrapolate in order to classify the whole SDSS. While the data and results we obtain are well-behaved, it will always be the case in astronomy that some form of extrap-

olation is ultimately required. This result is simply due to the fact that photometry will always be available several orders of magnitude fainter than spectroscopy, due to the physical difficulties in obtaining spectra of faint sources. Thus while our supervised learning represents a vital proof-of-concept over a whole terascale survey, ideally it should be extended with semi-supervised or unsupervised algorithms to fully explore the regions of parameter space that lie beyond the available training spectra. It is worth noting, however, that our training features, the object colors, are largely consistent beyond the limit of the spectroscopic training set.

- The data size is such that integrating the SQL database with D2K via JDBC is impractical, and the data must be stored as flat files. As database engines become more sophisticated, however, it could in the future become possible to offload partial or entire classification rules to a database engine. Doing so, however, would require supercomputing resources for the database engine, which results in an entirely new class of problems.

In moving to the petascale, further issues include:

- Conventional hardware, in the form of large clusters of multicore compute nodes, is scaleable to the petascale, however, field-programmable gate arrays (FPGAs), graphical processing units, and Cell processors may be more suited to many data mining tasks, due to their embarrassingly parallel nature. The LCDM group, in collaboration with the Innovative Systems Laboratory at NCSA, has demonstrated results on FPGAs using an SRC-6 MAPE system (Brunner, Kindratenko, & Myers 2007), which include running a kNN algorithm, although the implementation is not trivial because the algorithm must be rewritten.
- For many applications on the petascale, the performance becomes I/O limited. This is quantified by Bell, Gray, & Szalay (2006), who apply Amdahl's law (Amdahl 1967) that one byte of memory and one bit per second of I/O are required for each instruction per second, to predict that a petaflop-scale system will require one million disks at a bandwidth of 100 MB s<sup>-1</sup> per disk. They also state that data should be stored locally (i.e., not transferred over the internet), if the task requires less than 100,000 CPU cycles per byte of data. Many contemporary scientific applications are such that local storage is favored by over an order of magnitude.

**Acknowledgments.** The authors acknowledge support from NASA through grant 05-AISR05-0144.

## References

- Amdahl, G. 1967, in AFIPS Conference Proceedings, Vol. 30, 483–485
- Ball, N. M., Brunner, R. J., Myers, A. D., & Tchong, D. 2006, *ApJ*, 650, 497
- Ball, N. M., Brunner, R. J., Myers, A. D., Hepler, N. E., Alberts, S., Tchong, D., & Llorà, X. 2007, *ApJ*, 663, 774
- Bell, G., Gray, J. & Szalay, A. 2006, *IEEE Computer*, 39, 110
- Brunner, R. J., Kindratenko, V., & Myers, A., 2007, NASA technical report
- York, D. G. et al. 2000, *AJ*, 120, 1579